U.S. DEPARTMENT OF COMMERCE
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NATIONAL WEATHER SERVICE
SYSTEMS DEVELOPMENT OFFICE
TECHNIQUES DEVELOPMENT LABORATORY

TDL OFFICE NOTE 76-4

Improving the Bias in MOS Ceiling and Visibility Forecasts

Richard L. Crisci

March 1976

Improving the Bias in MOS Ceiling and Visibility Forecasts

by

Richard L. Crisci

INTRODUCTION

In November 1975, we computed comparative verification scores for TDL's automated 5-category ceiling and visibility guidance forecasts, NWS local forecasts of ceiling and visibility, and persistence forecasts of the same elements, for the period January 1 through March 31, 1975. The categories for the MOS forecasts are shown in Table 1. Percent correct (PC), NWS Matrix Score (MS), and category bias[1] were computed on forecasts for 92 U.S. stations; a total of approximately 5,000 forecasts of each element were verified. For the MOS system, forecast projections were 12, 18, and 24 hr from numerical model cycle time.

Table 1. Ceiling and visibility categories used for MOS 5-category forecasts.

Category	Ceiling (ft)	Visibility (mi)
1	$\leq$ 100	$\leq$ 3/8
2	200-400	1/2-7/8
3	500-900	1-2 1/2
4	1000-1900	3-4
5	$\geq$ 2000	$\geq$ 5

Our guidance forecasts were generated from cool season (Oct.-Mar.) equations described in Technical Procedures Bulletin No. 120 (NWS, 1974). Categorical forecasts were determined from the probability forecasts so as to maximize the MS.

In general, the MOS forecasts verified well. Although we did not outscore either persistence or the local forecasts for the 12-hr projection[2], we did better than persistence for the other two projections; the other local forecast projections were 15 and 21 hr so no other direct comparison with locals

---

[1] Bias is defined as the number of forecasts of a category divided by the number of observations of that category. A bias of 1 indicates unbiased forecasts of that category.

[2] Persistence has a decided advantage in the so-called 12-hr projection since the observation used for the "forecast" occurs two hours prior to verification time. For MOS equations containing surface observation predictors, the input observation is six hours earlier than verification time.

was possible. The shortcoming of all the MOS forecasts, however, was the unsatisfactory bias in the lower two categories. In most cases, the MOS system made very few forecasts in either category 1 or 2. Since these categories are of great concern to aviation interests, our product would not likely be considered useful. Therefore, we experimented with a transformation procedure to minimize the bias to see if we could achieve a better, more useful product. The purpose of this note is to describe our effort and the results we obtained.

PROCEDURE

We began with the dependent data sample which was used to derive the MOS prediction equations. These data consist of PE and trajectory model predictors from the cool seasons of 1969 through 1973--approximately 600 cases. We applied the data to our 5-category, regionalized equations for 233 terminals and produced forecasts of visibility for projections of 12, 18, and 24 hr from the 1200 GMT numerical model cycle. The forecasts were generated by the M700 program of the MOS development system (Glahn, 1973). We decided to restrict our effort to one forecast element for one cycle in order to keep the total time spent on the experiment to a minimum. We chose visibility in the 1200 GMT cycle because our bias scores were the poorest for that combination.

The next step was to use the forecasts, in conjunction with verifying observations, to determine threshold probabilities which would minimize the bias for the dependent data sample in transforming the probability forecasts to categorical forecasts. This determination was required for each of the lower four categories, for each projection, and for each of the 14 regions used in our 5-category MOS forecast system.

The threshold probabilities were developed in a cumulative fashion. Begining with the breakpoint for the lowest category, we computed the critical probability which would minimize the bias of that category for the dependent data. Next, we repeated the procedure with the breakpoint for category 2. This resulted in the determination of a probability which minimized the bias for all dependent data cases below that breakpoint, which included cases in category 1. We repeated this procedure two more times, with the breakpoints for categories 3 and 4. Because of the cumulative property of the computations, the threshold probability for each category is higher than that for its lower neighbor. An example of a set of threshold probabilities is shown in Table 2; the numbers given are those determined for region 7 (north-central U.S.) for the 18-hr projection.

Table 2. Threshold probabilities in percent determined to minimize the category bias of MOS 5-category visibility forecasts made from dependent data for the 18-hr projection, in region 7, 1200 GMT cycle.

	Category			
	1	2	3	4
Threshold probability	8	12	26	32

In practice, the threshold probabilities are used as follows: the forecast probabilities for the lower four categories are summed and compared to the threshold probability for category 4. If the sum does not exceed the threshold probability, category 5 is selected as the "best" category in transforming the probability forecast to a categorical forecast. If the sum does exceed the threshold probability, the forecast probability of category 4 is subtracted from the original sum and the new sum is compared to the threshold probability for category 3. If the new sum does not exceed the threshold probability, category 4 is chosen as the "best" category. If the sum does exceed the threshold probability, then the process just described continues until a "best" category is determined.

In the verification that we did which led to the experiment we're describing, we transformed the probability forecast by means of the NWS scoring matrix (NWS, 1973) shown in Table 3. It is designed to give credit to both hits and near-misses, with higher weights toward the lower, more difficult, forecast categories.

Table 3. NWS scoring matrix used to judge the usefulness of aviation forecasts.

Observed Category	Forecast Category				
	1	2	3	4	5
1	100	60	20	0	0
2	80	90	50	20	0
3	40	70	80	50	20
4	20	30	60	80	50
5	0	10	30	50	70

If the forecast probabilities for each category are designated as P_1, P_2, P_3, P_4, and P_5, and the weights in the matrix are designated as $W_{c,r}$ where c refers to the column and r to the row, an expected score (ES) for each category can be determined as follows:

$$ES_1 = P_1 (W_{1,1}) + P_2 (W_{1,2}) + P_3 (W_{1,3}) + P_4 (W_{1,4}) + P_5 (W_{1,5})$$

$$ES_2 = P_1 (W_{2,1}) + P_2 (W_{2,2}) + P_3 (W_{2,3}) + P_4 (W_{2,4}) + P_5 (W_{2,5})$$

and so on for all five expected scores. To transform the forecast, we chose the category yielding the largest ES. The same procedure is currently being used daily to select the "best" category for our operational forecast guidance product.

After we established appropriate threshold probabilities for the dependent data sample, we applied them in a reverification of the independent data of 1975. This time, we transformed the probability forecasts to categorical forecasts by using the threshold probabilities rather than the NWS scoring matrix. We then computed new scores for category bias, PC, and MS. While we were primarily intersted in improving the earlier bias scores, we were also concerned about degrading the other verification scores as a consequence. Therefore, we also needed to recompute PC and MS.

RESULTS AND CONCLUSIONS OF THE EXPERIMENT

In Table 4, we present the results of the new verification, along with the original results, for visibility forecasts in the 1200 GMT cycle for 12-, 18-, and 24-hr projections. We've also included the statistics for local 12-hr forecasts and persistence forecasts.

Table 4. Comparative verification of persistence, MOS 5-category, and local visibility forecasts, 1200 GMT cycle, for the period January-March 1975, for 92 stations. PC is percent correct, MS is NWS matrix score.

Projection	Type	Bias by Category					PC	MS
		1	2	3	4	5		
12 hr	MOS Original	.06	.11	.42	1.16	1.04	87.7	66.2
	MOS Latest	.13	.44	.77	1.01	1.02	87.1	66.1
	Persistence	.69	.98	1.04	1.01	1.01	89.1	67.3
	Local	.50	.76	.65	1.64	1.00	88.8	67.3
18 hr	MOS Original	.00	.08	.28	.94	1.07	86.1	65.2
	MOS Latest	.38	.98	1.00	1.11	1.00	83.5	64.7
	Persistence	.17	1.08	1.17	.89	1.01	83.7	64.5
24 hr	MOS Original	.02	.00	.29	.94	1.12	79.5	62.2
	MOS Latest	.50	.74	1.06	.85	1.03	76.5	61.7
	Persistence	.13	.67	.84	.68	1.08	76.2	60.9

Table 4 indicates we were successful in our attempt to improve the bias in the MOS 5-category visibility forecasts. In almost every instance, the biases in the "latest" verification are much closer to 1 than in the "original" and, for the 18- and 24-hr projections, the newer biases are better than those for peristence, which was not true for the earlier verification. For the 12-hr projection, the biases are improved but are still not as good as those for persistence for the lower three categories. As mentioned earlier, the 4-hr time advantage for persistence is extremely important. The other factor which had a strong effect on the results has to do with the determination of threshold probabilities for the lower two categories. The climatological frequencies of categories 1 and 2 are, in general, small--for the average terminal they're on the order of 1 to 2% which qualifies for rare event status. In our dependent data sample, therefore, we had relatively few cases to determine threshold probabilities with. For some regions--the far west and Florida--there were almost no cases at all. This meant that while the threshold probabilities we chose yielded a bias near 1.0 for the dependent data, a deviation in the threshold of one or two percentage points from that choice frequently made a drastic change in the bias. As an example, our choice of 10% for the threshold probability for category 1 in region 9 (eastern Texas and some southern states) for the 12-hr projection produced a bias of .98 on

the dependent data. However, a choice of 9% yielded a bias of 2.48. and 11% gave .02 for the same data. The obvious instability of the threshold probabilities could therefore be expected to have a major effect on the results for independent data. This would explain why the biases for category 1 in our reverification were not closer to 1.0.

We must also acknowledge the decrease in the PC and MS statistics, particularly the PC scores for the last two projections. The newer figures show we've lost the improvement we originally displayed over persistence. The reason for this change is basically the inaccuracy of the prediction equations. Improving the biases meant we made substantially more forecasts of the lower categories but, unfortunately, not many of those were hits. To illustrate this, Table 5 is a presentation of the contingency tables we compiled to compute the various verification scores shown in Table 4.

Table 5. Contingency tables used to compute verification scores shown in Table 3.

a. 12-hr projection

MOS guidance forecasts

Original

Observed Category	Forecast Category 1	2	3	4	5	T
1	0	1	8	1	6	16
2	0	1	9	12	32	54
3	1	3	51	72	159	286
4	0	0	20	50	152	222
5	0	1	32	123	4386	4542
T	1	6	120	258	4735	5120

Latest

Observed Category	Forecast Category 1	2	3	4	5	T
1	1	1	8	3	3	16
2	0	4	14	10	26	54
3	1	8	80	55	142	286
4	0	3	39	38	142	222
5	0	8	80	118	4336	4542
T	2	24	221	224	4649	5120

Persistence

Observed Category	Forecast Category 1	2	3	4	5	T
1	1	5	3	3	4	16
2	3	12	18	6	15	54
3	4	24	135	44	79	286
4	0	3	55	71	93	222
5	3	9	86	101	4343	4542
T	11	53	297	225	4534	5120

Local

Observed Category	Forecast Category 1	2	3	4	5	T
1	3	2	6	3	2	16
2	0	11	18	15	10	54
3	3	16	93	93	81	286
4	0	5	27	99	91	222
5	2	7	41	153	4339	4542
T	8	41	185	363	4523	5120

Table 5. continued:

b. 18-hr projection

MOS guidance forecasts

Original

Observed Category	Forecast Category 1	2	3	4	5	T
1	0	0	6	9	51	66
2	0	0	4	14	32	50
3	0	1	22	59	174	256
4	0	1	14	40	204	259
5	0	2	25	121	4374	4522
T	0	4	71	243	4835	5153

Latest

Observed Category	Forecast Category 1	2	3	4	5	T
1	2	7	11	6	40	66
2	1	1	15	12	21	50
3	5	13	65	43	130	256
4	7	8	44	45	155	259
5	10	20	120	181	4191	4522
T	25	49	255	287	4537	5153

Persistence

Observed Category	Forecast Category 1	2	3	4	5	T
1	1	1	15	11	38	66
2	0	3	9	9	29	50
3	2	16	68	39	131	256
4	3	6	49	41	160	259
5	5	28	159	130	4200	4522
T	11	54	300	230	4558	5153

c. 24-hr projection

MOS guidance forecasts

Original

Observed Category	Forecast Category 1	2	3	4	5	T
1	0	0	10	19	115	144
2	0	0	6	16	74	96
3	1	0	24	76	272	373
4	1	0	23	62	261	347
5	1	0	46	154	4092	4293
T	3	0	109	327	4814	5253

Latest

Observed Category	Forecast Category 1	2	3	4	5	T
1	11	11	22	10	90	144
2	4	2	23	15	52	96
3	13	14	87	60	199	373
4	12	13	73	43	206	347
5	32	31	189	166	3875	4293
T	72	71	394	294	4422	5253

Persistence

Observed Category	Forecast Category 1	2	3	4	5	T
1	2	2	22	11	107	144
2	0	3	16	13	64	96
3	3	10	49	38	273	373
4	2	5	44	35	261	347
5	11	44	183	140	3915	4293
T	18	64	314	237	4620	5253

In the 12-hr projection, we made only one new category 1 forecast, which was a hit, and 18 additional forecasts of category 2, of which only 3 were correct. For category 3, we made more than 100 new forecasts but less than 30 were correct. We made fewer forecasts for categories 4 and 5, of course, and lost a total of 62 previously correct forecasts in the process. The overall changes were relatively small, however, so the verification scores did not change very much.

Significant changes appeared in the reverification of the 18-hr projection. From originally making no forecasts of category 1, we then made 25 new forecasts and were correct in 2 cases. We fared even worse for category 2--45 additional forecasts and one hit! For category 3, we made 184 additional forecasts and were correct in 43 cases which was considerably better than our scores for the lower two categories. We continued to lose ground, with respect to our overall percent correct score, in category 4. There, we made 44 new forecasts and picked up only 5 hits. The total effect of these changes can easily be seen in the new statistics for category 5. We made 298 fewer forecasts of that category and gave up 183 previously correct choices in doing so. The new percent correct score for the entire sample shows we lost 2.6% as a consequence.

We could perform a similar analysis for the 24-hr projection, but the figures are analogous and the final figures are almost identical--the new percent correct score is 3% lower than the original. The point we want to make is that we need to develop prediction equations which are more skillful in predicting the lower categories so that we can achieve good biases without sacrificing the desired accuracy for our other measures of effectiveness.

All things considered, the results of our experiment were definitely mixed. We did improve the bias, but at the expense of both our other verification scores and our earlier improvement over persistence. Based on our findings, however, we will most likely use the concept of threshold probabilities to minimize the bias of our MOS prediction equations for ceiling and visibility. We will not develop threshold probabilities for our present warm season equations because there is not enough time to do so and meet implementation deadlines this year. We do expect to use threshold probabilities with the equations we are now developing, from LFM predictors, for implementation this fall.

ACKNOWLEDGMENTS

I would like to thank Robert Bermowitz for his help and guidance in deriving the threshold probabilities through the use of a computer program he developed for a similar need, George Hollenbaugh for his work to reprogram the relevant software, and Harry R. Glahn for his overall guidance and advice which helped to make our efforts successful.

REFERENCES

Glahn, Harry R., 1973: The TDL MOS development system, CDC 6600 version. TDL Office Note 73-5, 71 pp.

National Weather Service, 1974: The use of model output statistics for predicting ceiling and visibility. Technical Procedure Bulletin No. 120, 10 pp.

National Weather Service, 1973: Combined aviation/public weather forecast verification. Operations Manual Chapter C-73, 14 pp.

Table 5. Contingency tables used to compute verification scores shown in Table 3.

a. 12-hr projection

MOS guidance forecasts

Original

		Forecast Category					
		1	2	3	4	5	T
Observed Category	1	0	1	8	1	6	16
	2	0	1	9	12	32	54
	3	1	3	51	72	159	286
	4	0	0	20	50	152	222
	5	0	1	32	123	4386	4542
	T	1	6	120	258	4735	5120

Latest

		Forecast Category					
		1	2	3	4	5	T
Observed Category	1	1	1	8	3	3	1
	2	0	4	14	10	26	1
	3	1	8	80	55	142	2
	4	0	3	39	38	142	2
	5	0	8	80	118	4336	45
	T	2	24	221	224	4649	51

Persistence

		Forecast Category					
		1	2	3	4	5	T
Observed Category	1	1	5	3	3	4	16
	2	3	12	18	6	15	54
	3	4	24	135	44	79	286
	4	0	3	55	71	93	222
	5	3	9	86	101	4343	4542
	T	11	53	297	225	4534	5120

Local

		Forecast Category					
		1	2	3	4	5	
Observed Category	1	3	2	6	3	2	1
	2	0	11	18	15	10	5
	3	3	16	93	93	81	28
	4	0	5	27	99	91	22
	5	2	7	41	153	4339	454
	T	8	41	185	363	4523	512

Table 5. continued:

•b. 18-hr projection

MOS guidance forecasts

Original

Observed Category	Forecast Category 1	2	3	4	5	T
1	0	0	6	9	51	66
2	0	0	4	14	32	50
3	0	1	22	59	174	256
4	0	1	14	40	204	259
5	0	2	25	121	4374	4522
T	0	4	71	243	4835	5153

Latest

Observed Category	Forecast Category 1	2	3	4	5	T
1	2	7	11	6	40	66
2	1	1	15	12	21	50
3	5	13	65	43	130	256
4	7	8	44	45	155	259
5	10	20	120	181	4191	4522
T	25	49	255	289	4537	5153

Persistence

Observed Category	Forecast Category 1	2	3	4	5	T
1	1	1	15	11	38	66
2	0	3	9	9	29	50
3	2	16	68	39	131	256
4	3	6	49	41	160	259
5	5	28	159	130	4200	4522
T	11	54	300	230	4558	5153

Table 5. continued:

c. 24-hr projection

MOS guidance forecasts

Original

Observed Category	Forecast Category 1	2	3	4	5	T
1	0	0	10	19	115	144
2	0	0	6	16	74	96
3	1	0	24	76	272	373
4	1	0	23	62	261	347
5	1	0	46	154	4092	4293
T	3	0	109	327	4814	5253

Latest

Observed Category	Forecast Category 1	2	3	4	5	T
1	11	11	22	10	90	144
2	4	2	23	15	52	96
3	13	14	87	60	199	373
4	12	13	73	43	206	347
5	32	31	189	166	3875	4293
T	72	71	394	294	4422	5253

Persistence

Observed Category	Forecast Category 1	2	3	4	5	T
1	2	2	22	11	107	144
2	0	3	16	13	64	96
3	3	10	49	38	273	373
4	2	5	44	35	261	347
5	11	44	183	140	3915	4293
T	18	64	314	237	4620	5253